IS

Lectures (1 to 6)

IS

* Skills of data scientist

1) statistices skills. 2) Database.
3) Critical thinking, creative, Adaptive, communication skills.
4) machine learning + Data mining + Advanced mathematics.
5) Collect data from different online source.
6) Extract data & Analysis 7) Programming skills.
8) web (development + design).
9) Can make correlations & connections.

---

* Data enables Proffessional (data collectors)
↳ Quantitive (can measure upcoming data & give technical reports on it)
↳ skeptical (be
↳ Communications & collaborative.

Big data

↳ refers to the exponential growth and availability of data, both structured and unstructured.

* Three V's describe definition of big data?

1) Volume 2) velocity 3) variety

1 Volume

↳ There is a large increase of data volume (why)?

a. all of transactional data that has been added up over the years.

b. streaming data from social media.

c. machine to machine data increase.

← التحدى الكبير إزاى نتحرك بسرعة مع ال (data) السريعة.

2 velocity

↳ Data is being streamed at huge speeds and need to be dealt with any timely manners like (social media & Mobile devices)

3 Variety

↳ many different of data

a. Email

b. Numeric data

c. structured data

d. unstructured documents

e. Audio & video

f. Application data.

← منظمات كتيره جداً بتصارع على انها تتحكم فى أنواع ال (data) المختلفة.

Veracity مزود 4 V's لو قالك انهم

* Veracity (uncertainty of data)

↳ refers to the trustworthiness of the data.

with many forms of big data (quality & accuracy)

are less controllable.

5 V's ⇒ Value of data is added

↳ well and good for access or useless data

Big data

↳ data whose scale, distribution, diversity and/or timeless require the use of technical architectures and analytics to enable.

* Key characteristics of Big data.

1) data volume.

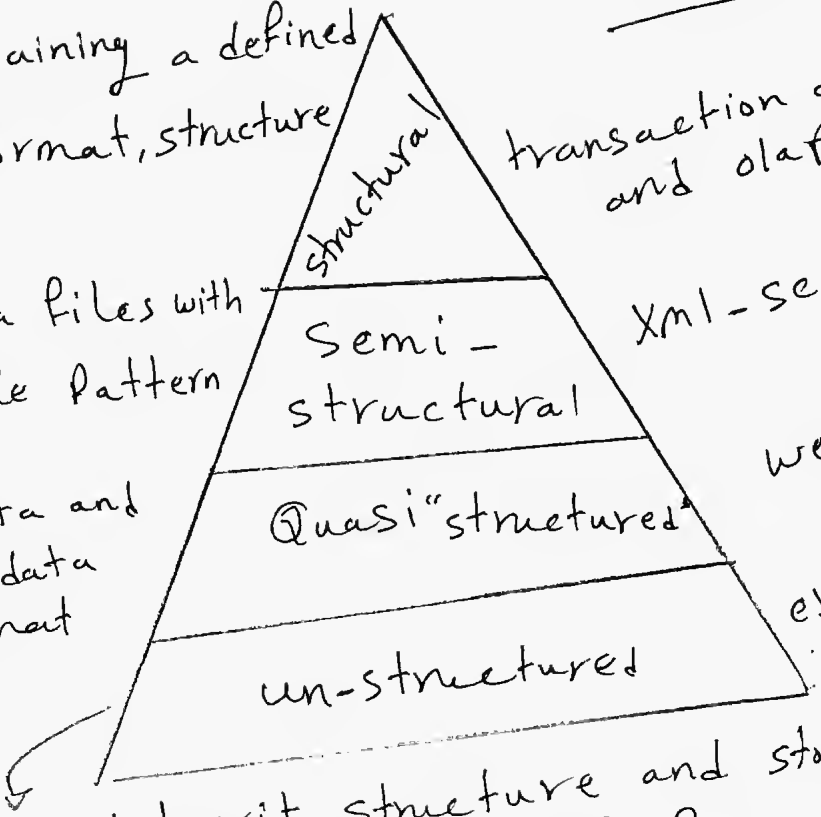
2) Processing complexity

3) Data structured.

examples

Data containing a defined datatype, format, structure

transaction data and olap



textual data files with a discernable pattern

xml-schema

textual data and unorganized data format

web click stream

ex: pdf

data has no inherit structure and stored as different types of files.

Data islands - spread ~~sheets~~ marts	Data warehouse	Analytic sandbox
isolated data.	centralized data containers in a purpose-built space	Data assets gathered from multiple sources & technologies for analysis
Analyst dependent on data extracts.	Analyst dependent on IT & DBAs for data access and schema changes	analyst owned gives high performance reduce cost associated.

Business intelligence	Data science
→ structured data, traditional sources, manageable datasets.	→ structured / unstructured data & multiple types of sources & very large data sets
standard	optimization, predictive modeling, statistical analysis.
his questions did How many [redacted] we sell ? where is the Problem?	What if ...? openended questions?

* criteria of Big Data Projects

1) Speed of decision making.

2) Throughput. 3) Analysis flexibility.

* Data scientist Key Activities

1) reframe business challenges as analytical challenges.

2) Design &, implement and deploy statistical models and data mining techniques of big data.

3) create insights that lead to actionable recommendations.

* Three Keys roles of the new data Ecosystem:-

1) Deep analytical talent → People with advanced training in quantitative disciplines such as math, statistics, machine Learning.
↳ Data science

2) Data savvy Professionals → People with basic Knowledge of statistics and/or machine learning who can define Key questions that can be answered using advanced analytics.

3) Technology & data enablers → people Providing technical expertise to support analytics Projects skill sets including computer Programming & DB adminstrator.

* Key roles for successful Analytic Projects:-

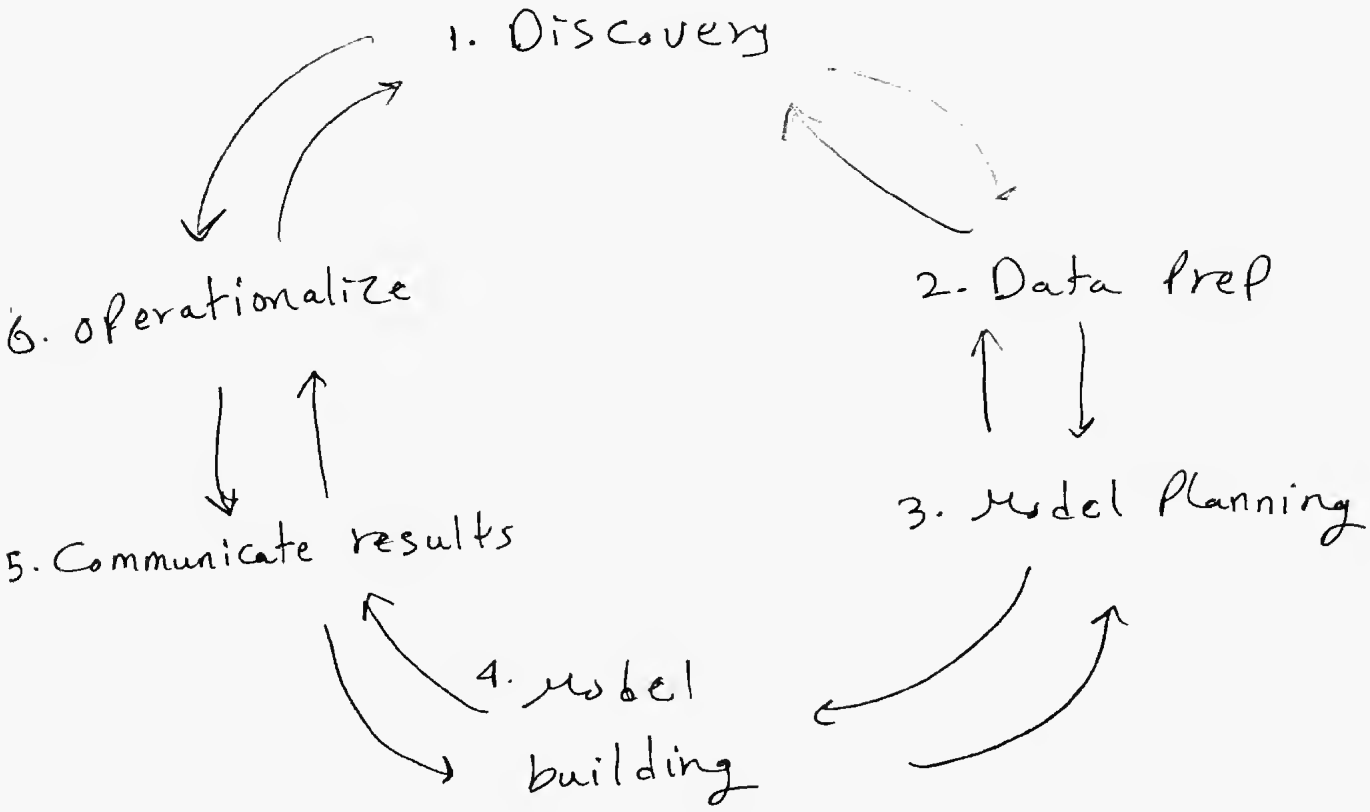
Role.	description
Bussiness user	→ benefits from end results, Can Consult and advise Project team on Value of end results.
Project manager	→ He care only about output → ensure Key objectives are met on time and at expected quality.

	description
Project sponser	→ Provide the fund needed in Project. → Cares only about completed work. → responsible for Genesis of projects.
Business intelligence Analyst	→ measure the indicators from point of view of business
Data engineer	→ responsible for data (with its variations) → extract data → has deep techincal skills for data management.
Data administrator (DBA)	→ Provisions and configures database environment to support analytical needs of working team.
Data scientist	→ deal with data analytically. → ensure that overall analytically objectives are met.

* Value of using data analytics lifecycle?

1) ensure rigidity and completeness.
2) enable better transition to members of the cross-function analytic teams.

Data analytics life cycle



- لو وصلت لمرحلة ولقيت ان ال (data) اللي معاك غير كافية ترجع للخطوة السابقة وتعيد حساباتك.

Data Analytics lifecycle

- reverse & forward لـ شغلانة

① Discovery

لـ بعرف بعد اذل خطوة هلاقيها (data) كانية
اعمل بيها (analytic plan) لو فيه حاجات ناقصه
يبقى الخطوه لسه مخلصتش.

② Data Prep. (Collect data)

بعدها
لـ هلاقيها (data) كانية اعمل بيها (build for model)

③ Model Planning

لـ بعدها هعمل (planning) حوا على أساس صحيح
لـ لو تمام هكمل

④ Model build

لـ اعمل (build) وبعدين هعمل (test for model)

⑤ Communicate results

لـ لازم يكون عندك (skills) تقنع بيها غيرك بالنتايج
اللى حققتها.

⑥ operationalize

1 Discovery

↳ Problem defination phase.

↳ need to learn about domain we are working on

↳ Know more about history of this domain

له بشوف اللى قبلى عملوا ايه و وصلوا لفين و ايه المشاكل اللى واجهتهم وهل التجربة نجحت ولا لأ.

له هل صرعليا (Project) قبلي ولا..

له لازم أقيس ال (resources) اللي معايا.

---

2 Data Preparation

له بعمل (Prepare) لل (data)

له لل (Phase) اللى بتعمل فيها (sandbox)

← شوف مدرستين

a) ELT ← extract - load - transform

b) ETL ← extract - transform - load

~~extract data from~~ dataWarehouse

↳ load data in sandbox.

↳ transform (لو عندك (data) فى مختلفة (scales)

له ده ضيط الشغل بتاعك

له بعمل (transform) فى (range) واحد.

~~Conversion التحويل (transform) لـ~~

ال (2 Phases) أكثر (Preparation ، Discovery)

بتحويل فيهم (refine) &

3 Model Planning

لـ بناءً على اللي حققوه 1، 2 يبدأ أحدد الطريقة بتاعي هيمشي إزاي.

لـ بعمل (Feature selection)

لـ بفكر وأربط الـ (Work) اللي criteria هيمشي عليها.

4 Model build

لـ الـ (implementation)

لـ بعمل (test) لـ (model) بتاعي

5 Communicate results

لـ بعرض النتائج بتاعتي بالـ (benefits) اللي وصلت ليها.

6 Operationalize

لـ بعمل (operation) للشغل بتاعي.

* 4 Core deliverables to meet most stakeholders needs :-

[1] Presentation for Project Sponsers:-

- Big picture take aways for executive level stakholders
- determine key messages to aid their decision-making process.
- focus on clean, easy visuals for Presenter to explain & for viewer to grasp.

[2] Presentation for analysis:-

- Business Process changes.
- reporting changes.
- data scientists want the details.

[3] Code : for technical people

[4] Technical specs: of implementing the code.

* Analyst wishlist for a successful analytics Project:-

[i] Data & workspaces

a. access all data. b. sandbox

c. Ability to move data back between staging.

d. up-to-date data dictionary.

2 Tools

a. statistical, matematical, visual SW.

b. tool or place to log errors with systems.

c. Collaboration → online platform for communication between team members.

Sandbox: Data assets gathered from multiple sources and technologies for analysis.

↳ high performance analytics.

↳ reduce costs of data replication.

↳ Analyst owned.

Tools used in lifecycle

1 Data Prepration

- descriptive statistics.
- visualization (R), spotfire

→ for data transformation

↳ sql, Hadoop, Mapreduce.

[2] Model Planning

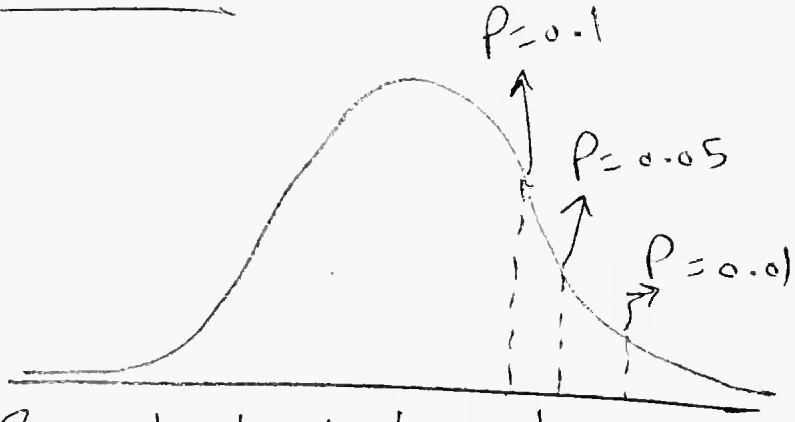
- R / Postgres SQL, sql analytics, Apline miner, spss / oBDc.

[3] Model Building

↳ R, PL/R, SQL, ~~SAS/ACT~~

---

*Distribution of sample means



→ Calculate the mean

$$\frac{\Sigma(x - \bar{X})}{N}$$

→ calculate variance & standard deviation

$$s^2 = \frac{\Sigma(x - \bar{X})^2}{n} \quad \& \quad s$$

→ calculate the p value

if p-value is between 0.01 & 0.1
↳ inside range (normal case)

if not → reject ~~the~~ null hypothesis.

Notes

*significance
↳ Probability of False Positive (α)

*Power
↳ Probability of a true Positive ($1-\beta$)

*Effect size
↳ size of observed difference.